GINSENG : une grille dédiée à l'e-santé et l'épidémiologie

Paul de Vlieger, Sylvain Planche, David Manset, Jérôme Revillard, David Sarramia, Lydia Maigne

GINSENG : une grille dédiée à l'e-santé et l'épidémiologie

P. De Vlieger(1), S. Planche(2), D. Manset(3), J. Revillard (4), D. Sarramia (5) and L. Maigne (6)

(1) (2) (5) (6) [vlieger|planche|maigne][sarramia]@clermont.in2p3.fr, Clermont Université, Université Blaise Pascal - LPC CNRS/IN2P3 BP 10448, F-63000 Clermont-Ferrand
(3) (4) [dmanset|jrevillard]@maatg.fr, MAAT-France, 74070 Archamps, France

Overview

Emerging challenge concerning public health statistics is the ability to provide real time information on population health. It is especially relevant in case of emergency scenarios: pollution through toxic gas emission or radioactivity, heat waves, pandemic flu viruses. The daily improvement of care practice can also benefit of any real time information on patients hosted in medical structures.
To face this problematic, the french GINSENG project uses the european grid technology to create a sentinel network for e-health and epidemiology. This distributed network architecture offers many advantages:

- Medical data banks from each hospital or labs can be interrogated directly without centralizing any information
- Such architecture is then really cost effective.
- Statistical studies will be soon available in real time through a web interface accessible by the medical staff.

While patient data consistency can mainly be achieved by working on medical databases standardization, patient identification and medical data linkage mechanisms are performed dynamically through the grid network. Authentication and data encryption are ensured by healthcare professional smartcards containing an X509 grid-compatible certificate delivered by a trusted certification authority.
The GINSENG project focuses on two fields: cancer surveillance and perinatal health.

Enjeux scientifiques

Le projet ANR GINSENG vise à créer une infrastructure de grille dédiée au partage de données et aux études épidémiologiques pour la région Auvergne. Ce projet fédère différents acteurs :

- le Laboratoire de Physique Corpusculaire pour son expertise dans la gestion de bases de données médicales réparties ;
- le CHU de Clermont-Ferrand avec son service de biostatistiques chargé d'évaluer un nouveau type de signal en veille sanitaire et la mise en place d'un système opérationnel de détection et d'alerte dans la surveillance de certaines pathologies et avec son service de santé publique chargé de d'analyser si ce nouveau type d'architecture distribuée permet une amélioration significative des pratiques médicales ;
- enfin la société maatG France pour son expertise dans le développement d'interface conviviale aux grilles informatiques.

Dans un premier temps, le projet GINSENG va concerner deux applications médicales : le Réseau Sentinelle Cancer Auvergne (RSCA) et le Réseau de Santé Périnatale en

Auvergne (RSPA).

Le projet GINSENG permet d'envisager des innovations importantes dans deux champs majeurs de l'épidémiologie : la veille sanitaire et l'évaluation des politiques de santé. La veille sanitaire est l'un des enjeux primordiaux de santé publique pour les années à venir : l'accentuation des circulations humaines sur tout le globe se traduit par un risque accru de maladies émergentes. De la même manière, le développement des politiques de santé, et plus particulièrement des politiques de prévention secondaires, rend nécessaire de disposer d'outils adaptés à leur évaluation. Dans les deux cas, les mesures de risque ou d'efficacité se font jusqu'à ce jour à partir de recueils créés ad hoc avec toutes leurs limites : perte d'information, biais de sous déclaration, absence de données pour un risque non connu, biais de mesure (par exemple pour les données de nature médico-économiques).
Les technologies proposées dans ce projet doivent rendre possible un partage électronique et sécurisé de ces données de manière à les rendre disponible à tout instant dans le cadre notamment d'une veille sanitaire ou d'analyses épidémiologiques (allant de l'observation sanitaire à l'évaluation de prises en charge ou de politiques de santé). Le projet GINSENG permet de répondre à ce contexte sans nécessité de déclaration médicale mais en interconnectant des bases de données existantes et distribuées.
Les enjeux majeurs auxquels nous devons répondre sont :

- le coût de l'infrastructure ;
- le travail sur des données qui reposent sur le maintien de leur distribution et non leur centralisation ;
- l'interopérabilité ;
- les alarmes sanitaires efficaces, nécessitant moins d'étapes intermédiaires de déclaration et de traitement des données.

Développements, déploiement sur la grille

Architecture

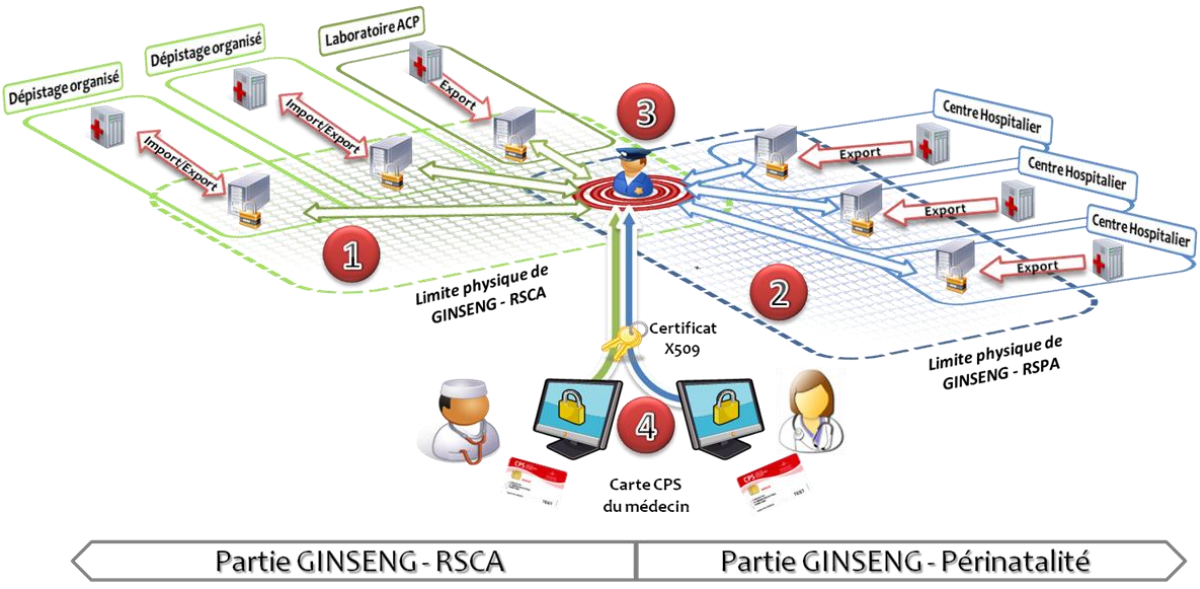

Figure 1 : Architecture de la grille GINSENG

L'architecture de grille mise en œuvre à ce jour (voir Figure 1) regroupe différents

acteurs:

- 1°) et 2°) **Les structures médicales** (associations de dépistage, laboratoires de pathologie et centre hospitaliers) : chacune de ces structures exporte automatiquement ses données depuis son serveur d'application vers le serveur de grille. Elle a aussi la possibilité d'importer les données du réseau.
  La manipulation des données et leur mise en réseau est gérée au niveau du serveur de grille par Pandora Gateway, une plateforme développée par la société maatG dans le cadre d'un projet pionnier dans le domaine des grilles pour la santé : Health-e-Child [1]. Cette plateforme offre une couche d'abstraction de la complexité des grilles et permet, au moyen d'une architecture orientée services, de répondre à un grand nombre de problématiques de haut niveau pour la santé.
- 3°) **Les services centraux** : ce sont les services utilisés pour la gestion du système d'information et de la sécurité. Une plateforme Web permet d'offrir une interface pour l'accès sécurisé aux données.
- 4°) **Les utilisateurs du système** (médecins épidémiologistes, associations de dépistage): possesseurs de la carte de professionnel de santé (CPS), ils peuvent s'authentifier et effectuer des requêtes statistiques dans leur domaine d'étude.

Intégration des données médicales

Toutes les données médicales présentes sur le réseau implémentent le modèle de données fourni avec Pandora Gateway: Federated Electronic Health Records (FedEHR). Ce modèle permet de décrire les données médicales de manière générique L'uniformisation offerte par FedEHR permet aux services d'interroger de manière transparente l'ensemble des sources de données.
FedEHR fournit un modèle constitué d'entités abstraites communes à tous les domaines médicaux : le patient, la visite, l'évènement médical et la variable clinique. Ces quatre entités sont organisées hiérarchiquement de manière à pouvoir représenter un historique médical : chaque patient effectue des visites ; chaque visite associe différents évènements médicaux ; chaque évènement médical est un ensemble de données cliniques. Aussi, FedEHR utilise un ensemble d'archétypes cliniques issus d'openEHR pour la gestion des types de données (Mesure, Observation, Annotation, Image, ...) qui permettent de couvrir les différentes types d'informations reçus des sources du réseau.
L'importation des sources de données se fait par l'implémentation d'un modèle composé d'éléments de FedEHR. La création de métadonnées associées aux éléments permet d'en décrire le contenu. De plus, l'utilisation des différents types de données cliniques permet de couvrir l'ensemble des types d'informations de la source. Enfin, le développement d'un programme d'importation permet de lire les données de la source et de les importer dans le modèle de FedEHR en associant celles-ci aux métadonnées précédemment créées.

Utilisation de la carte de professionnel de santé (CPS)

Les recommandations ASIP-Santé [5] imposent en France l'utilisation des cartes de professionnel de santé pour l'authentification des utilisateurs médicaux. Un module est implémenté pour coupler le système d'authentification de Pandora Gateway avec les API fournies par le GIP-CPS. Les autorités de certification délivrant les cartes CPS ont aussi été ajoutées dans VOMS pour vérifier leur validité.

Outils, difficultés rencontrées

Les outils utilisés proviennent du développement des grilles pour la santé [3]. Premièrement, VOMS[1] [4], constitue la charnière principale de toute la sécurité : en utilisant une infrastructure à clé publique, VOMS permet d'authentifier fortement les utilisateurs grâce à leur certificat électronique émanant d'une autorité de certification (AC) valide sur la grille. La création d'une organisation virtuelle pour l'ensemble des applications du réseau permet de centraliser l'administration des droits d'accès et de mutualiser les moyens. La séparation entre les réseaux se fait en définissant un groupe d'utilisateur par domaine d'application : RSCA et RSPA.
Ensuite, la mise en relation des données des différents serveurs des sites médicaux se fait grâce au moyen d'AMGA [6], un système de gestion de bases de données conçu pour les environnements de grille.
AMGA offre, en plus d'une base de données conventionnelle, des outils de réplication et de collaboration entre sites. De cette manière, les sites interconnectés implémentent le même schéma générique (FedEHR) qui est fédéré au niveau des services centraux de la grille. Les requêtes sont alors distribuées de façon complètement transparente pour les besoins de l'épidémiologie.
Une demande d'autorisation a été déposée à la CNIL, comprenant, avec la description du réseau et de ses objectifs, un ensemble de garanties sur la sécurité des données du patient et sur la possibilité qu'il puisse accéder, modifier ou s'opposer à leur traitement.

Résultats scientifiques

Des développements sont nécessaires afin d'assurer une bonne qualité des données à destination de l'épidémiologie. Plus particulièrement sur la question prépondérante de l'identification du patient. Il n'est pas possible pour l'instant, selon la loi informatique et libertés, pour des raisons de confidentialité, d'utiliser le numéro de sécurité sociale (NIR) pour le croisement de fichiers de santé. La question posée est simple : Comment assurer au sein d'une architecture distribuée l'identification d'un patient ? De cette interrogation en découle une seconde : Comment rapprocher des données patient réparties sans numéro commun ?

Modèle d'identification distribué et dynamique du patient

En attendant la généralisation de l'INS (Identifiant National de Santé) proposé par l'ASIP [5], il est nécessaire de recourir à d'autres méthodes de rapprochement de l'identité. L'identification d'un patient est une chose complexe : à l'heure actuelle, chaque établissement de santé stocke un numéro qui lui est propre. En regroupant ces différents identifiants internes au sein d'un nouvel identifiant unique et anonyme, du type UUID [7] et en les couplant avec des méthodes de chiffrement, l'identité du patient est assurée, et ceci de façon sécurisée au sein de l'ensemble du réseau. Cette méthode possède l'avantage d'être très souple. En effet, le simple ajout d'un identifiant à une liste agrégée permet de prendre en charge un nouvel établissement de santé. Elle permet

[1] Virtual Organisation Management System

aussi simplement de gérer le rapprochement comme la séparation d'identités.

Rapprochement d'identités patient réparties sans identifiant commun

Cette problématique, plus communément appelée "Record Linkage" ou "Medical Record Linkage" [8] est un procédé issu de la fouille de données visant à se servir de toutes les informations à disposition afin de rapprocher des sources hétérogènes. Cette technique est nécessaire lorsque les sources de données à rapprocher peuvent contenir des biais comme il est fréquent de l'observer d'un établissement de santé à l'autre : fautes de frappe, erreur de retranscription orale, changement d'adresse, etc. Une méthode combinant un algorithme analytique : Jaro-Winkler [8] et phonétique : Phonex [9] a été proposée. Elle permet, en combinant les noms, prénoms, date de naissance et adresse d'obtenir un taux de rapprochement de 95% sans erreur d'identification. Ainsi, une grande partie des données sont clairement identifiées et peuvent directement être utilisées à des fins épidémiologiques.

Perspectives

L'architecture de grille mise en œuvre dans le cadre du projet GINSENG regroupe déjà les principaux acteurs du dépistage organisé des cancers dans la région Auvergne. En attendant les autorisations CNIL nécessaires, les développements informatiques concernant l'intégration des métadonnées sur le réseau seront testés avec des bases de données médicales simulées à partir des données réelles recueillies. La phase expérimentale du projet sur le dépistage des cancers permet de récupérer 30 dossiers par semaine. Les bases de données de RSPA sont quant à elles plus volumineuses et nécessitent un traitement plus poussé pour gérer les métadonnées associées. A terme, une fois cet effort d'intégration terminé, une dizaine de maternités de la région seront équipées d'un serveur de grille permettant ainsi une interrogation des bases à distance. Les premiers indicateurs épidémiologiques issus de l'interrogation des bases de données médicales seront mis en place dès la fin de l'année 2011. L'évaluation des compromis réactivité/spécificité du système d'alerte sera comparée à celle des pratiques médicales déjà existantes.

Références

[1] J. FREUND, ET AL., *Health-e-child: an integrated biomedical platform for grid-based paediatric applications.* Stud Health Technol Inform, 2006. **120**: p. 259-70.

[2] P. DE VLIEGER, ET AL., *Grid-enabled sentinel network for cancer surveillance.* Stud Health Technol Inform, 2009. **147**: p. 289-94.

[3] V. BRETON, K. DEAN, ET T. SOLOMONIDES, *The Healthgrid white paper.* Studies in health technology and informatics, 2005. **112**: p. 249.

[4] R. ALFIERI, ET AL., *From gridmap-file to VOMS: managing authorization in a Grid environment.* Future Generation Computer Systems, 2005. **21**: p. 549-558.

[5] ASIP-SANTE. *Agence des Systèmes d'Information de santé Partagés*. Disponible à: http://esante.gouv.fr

[6] B. KOBLITZ, N. SANTOS, ET V. POSE, *The amga metadata service.* Journal of Grid

Computing, 2008. **6**(1): p. 61-76.
[7] P. LEACH, M. MEALLING, ET R. SALZ, *A Universally Unique IDentifier (UUID) URN Namespace*. 2005, RFC 4122.
[8] W.E. WINKLER, *Overview of record linkage and current research directions.* Research Report Series, 2006. **2**.
[9] A. LAIT ET B. RANDELL, *An assessment of name matching algorithms.* Technical Report Series - University of Newcastle Upon Tyne, 1996.